ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 5, May 2015

Twitter Event Summarization Using Phrase Reinforcement Algorithm and NLP Features

Mr. Ganesh Mane[1], Mrs. Anita Kulkarni[2]

Research Scholar, Computer Science & Engineering Department, Walchand Institute of Technology, Solapur, India[1]

Assistant Professor, Computer Science & Engineering Department, Walchand Institute of Technology, Solapur[2]

Abstract: Now a day's social networking sites are the fastest medium which delivers news to the user as compared to the newspaper and television. There so many social networking sites are present and one of them is Twitter. Twitter allows large no. of users to share/post their views, ideas on any particular event. According to recent survey, daily 340 million Tweets are sent on Twitter which is on a different topic and only 4% of posts on Twitter have relevant news data. It is not possible for any human to read the posts to get meaningful information related to specific events. There is one solution to this problem, i.e. we have to apply Summarization technique on it. In this paper, we have used an algorithm which uses a frequency count technique along with this we have also used some NLP features to summarize the event specified by the user. This automatic summarization algorithm handles the numerous, short, dissimilar, and noisy nature of tweets. We believe our novel approach helps users as well as researchers.

Keywords: Phrase Reinforcement Algorithm (PRA), Twitter API, Twitter, Natural Language Processing (NLP), Textual Entailment, Word Sense Disambiguation, WordNet, ROUGE Toolkit.

I. INTRODUCTION

Now a day's a social networking site plays the important role in human life such as Twitter. Twitter gives more information of current world events such as the earthquake in Nepal. With the help of Twitter, a user can create and share ideas and information instantly, without any barrier. But there is one problem daily 500 million Tweets are sent on Twitter and they are not in serial order. Summarization of Twitter events helps to fight with this problem.

Tweets under a Trending Topic contain a wide variety of useful information from many perspectives about important events taking place in the world. Basically, a summary that provides representative information of topics with no redundancy and well-written sentences would be preferred.

According to recent survey[13] daily 500 million Tweets are sent on Twitter which is on a different topic such as Sport related event, Tweets by News agencies, Business news, Tweets posted celebrity etc. It is difficult to find Tweets related to a particular topic. It is not possible for humans to read each and every Tweet to get correct and accurate information related to a specified event.

The texts from News articles, Books, Research paper, etc. are usually formal writings and have the highest language quality. On the other hand, the language of tweets is highly noisy, spelling and grammar mistakes. Tweets contain Typos, abbreviations, phonetic substitutions, ungrammatical structures, and emoticons, etc. Due to the above characteristics, text summarization techniques, in general, may not adapt well to the Twitter text.

Following are some problem the user has to deal with while reading Tweets about any event:

1. 1] Language issue
2. 2] Short and noisy Tweets
3. 3] Some of the users are trying to divert the event
4. 4] Some of the tweets contain the only link.
5. 5] Spam-tweets

One solution to these problems is Twitter API. Twitter allows the only authorized user to access and use functionality provided by the Twitter API. Twitter API allows the user to retrieve Tweets related to the specified topic, event etc... It also provides some functions to filter the Tweets such as removing non-English Tweets, getting tweets related to the specific event, etc. Spam-tweets can be easily removed since they almost always have a URL in them. The Twitter API allows a user to get only 100 tweets per day. But only 100 tweets are not sufficient to summarize any event and to collect more tweets we have to wait for some days. Instead of this we can use Twitter Stream API to download tweets related to specific events because Twitter Stream API allows downloading infinite no. of tweets. If there is a huge

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 5, May 2015

amount of tweets are available, then helps to generate a better summary.

II. LITERATURE REVIEW

Xiaobin Li, Stan Szpakowicz and Stan Matwin[10] has presented A WordNet based Algorithm for Word Sense Disambiguation. In this paper, they proposed an algorithm for automatic word sense disambiguation based on the lexical knowledge contained in Word Net and on the results of the surface-syntactic analysis. The algorithm is designed to support text analysis with minimal pre-coded knowledge, although the algorithm is assumed to aim at word sense disambiguation of noun objects in a text; in fact, it can be easily transformed to cover some other parts of speech in a text. Their approach focuses on two parts the full utilization of the important relationships between words in WordNet and the exploration of WSD heuristic rules based on the semantic similarity between words.

Joel Judd and Jugal Kalita[7] has presented the Better Twitter Summaries. In this technique, they are trying to improve the summary produced by PRA. The idea behind creating the desired summary is to parse the "raw" summary and build dependencies between the dependent and governor words in each summary. We perform parts of speech tagging and obtain lists of governing and dependent words. The Stanford Core NLP parser was used to build the lists of the governor and dependent words. They show the PR Algorithm can be improved by taking into account governor-dependency relationships among the constituents.

Hassan Sayyadi, Matthew Hurst and Alexey Maykov[5] has presented Event Detection and Tracking in Social Streams. In this paper, they propose a new algorithm for event detection using the co-occurrence of keywords. In our community detection algorithm, nodes can fall into different communities as a word or phrase can be in keywords list of more than one event. In the current version of our algorithm, we count all keywords in one community as keywords for the event, though a subset of keywords may be better, especially in cases where the number of nodes is large.

Gulab R. Shaikh and Digambar M. Padulkar[3] has presented Template Based Abstractive Summarization of Twitter Topic with Speech Act. In this paper they proposed work; the speech act-guided summarization approach is used to generate a summary of the twitter trending topic. With the recognized speech acts, the next step is to extract keywords and phrases from tweets to generate abstractive summaries. The extracted key terms are then ranked and inserted into special summary templates designed for speech acts by using the n-gram selection algorithm. This system is designed to accommodate the numerous, short, dissimilar, and noisy nature of the tweets. This proposed approach makes a good work contribution to the summarization community.

Prodromos Malak's Otis and Ion Androutsopoulos[9] has presented Learning Textual Entailment using SVMs and String Similarity Measures. They proposed a textual entailment recognition system that relies on SVM s whose features correspond to string similarity measures applied to the lexical and shallow syntactic level. They have suggested two additional possible improvements: applying partial matching to all of the string pairs and investigating other feature selection schemes. In future work, they also plan to exploit WordNet to capture synonyms, hypernyms, etc.

Chin-Yew Lin[8] has presented ROUGE: A Package for Automatic Evaluation of Summaries. ROUGE stands for Recall-Oriented Understudy for Gisting Evaluation. In this paper, they introduced ROUGE, an automatic evaluation package for summarization and conducted comprehensive evaluations of the automatic measures included in the ROUGE package using three years of DUC data. This paper introduces four different ROUGE measures: ROUGE -N, ROUGE-L, ROUGE-W, and ROUGE–S. They had shown how to achieve a high correlation with human judgments in multi-document summarization.

III. METHODOLOGY

Phrase Reinforcement Algorithm

The Phase Reinforcement Algorithm works as follows. The algorithm starts with a starting phrase, which is the event for which one desires to generate a summary. Given the starting phrase, the PR algorithm submits a query to Twitter.com for a list of posts that contain the phrase. Given the returned set of posts, the algorithm next filters the posts to remove any spam or irrelevant posts.

To avoid unnecessary processing we have to apply filtration technique on Twitter Dataset to remove it. In this, we first remove those tweets which contain bad words. We remove any non-English Tweets, short tweets, and Tweet contains the only link. After this, we have to remove stop words from Tweets Dataset.

Once we have a relevant set of Tweets (Training posts), the PR algorithm formally begins. The central idea of the PR algorithm is to build an ordered acyclic graph of all the words within the set of training posts. While constructing graph we check that whether that word is already present in the graph or not. If the word is already

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 5, May 2015

present we increment the weight of that word otherwise we create a new with initial weight one.

Phrase Reinforcement Algorithm[2] is referred from the paper Automatic Summarization of Twitter Topics and added two NLP features in it.

Following two Natural Language Processing (NLP) features we are using along with the algorithm:

Word sense disambiguation[10] (WSD) is the ability to identify the meaning of words in context. Given a set of words (e.g., a sentence or a bag of words), a technique is applied, which makes use of one or more sources of knowledge to associate the most appropriate senses with words in context. To determine the sense of words we can use machine-readable dictionaries, semantic networks. To implement word sense disambiguation we are using Word Net lexical database[11][12].

Consider the following sentences:
a) His speech was both witty and informative.
b) His speech was amusing.
In above two sentences, the word witty and amusing is synonyms mean both words have same meaning. We can use this feature while assigning a weight to the words. Textual entailment[9] (TE) challenge, which focuses on detecting semantic inference, has attracted a lot of attention. Given a text T (sentences) and a hypothesis H (one sentence), the goal is to detect if H can be inferred from T.

Consider the following sentences:
T: Phillip Hughes funeral: Watch Australian captain Michael Clarke's tearful tribute to batsman.
H: DailyMirror: Phillip Hughes funeral: Watch Australian captain Michael Clarke's tearful tribute to batsman. The above two Tweets has more than 90% similarity between them so we have to remove one tweet.

Collect the tweets from Twitter database for the event specified by user and store them into a file.

1) After collecting tweets apply filtration techniques, i.e. removes tweets which contain bad words, remove duplicate tweets, spam tweets and remove stop words.
2) Then count the frequency of each word form set of tweets by using Word sense disambiguation.
3) Get top ten words from frequency count and take only those tweets which contain one of the word from those ten words and transfer only those tweets for further processing.
4) Once again repeat step 3and 4.
5) Then apply Textual Entailment technique on the summary given by PR algorithm.

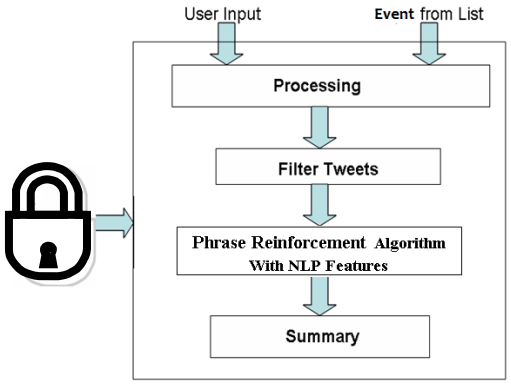

Figure. Conceptual Overview of Proposed System

In the first step, we take input from the user, i.e. Twitter event name if the user knows the event name and if not, then the user can select can event from the list. The event list contains top 10 current trends from Twitter for the selected region. After selecting input, we collect tweets from the Twitter site by using Twitter Stream API and store them into a file for further use.

In the second step, we apply filtration technique on collected tweets for removing Duplicate and noisy, irrelevant tweets, Spam-tweets, tweets which contain bad words, etc. After this, we remove stop words from the filtered tweets to apply Phrase Reinforcement algorithm (PRA).

In the third step, we apply the Phrase Reinforcement algorithm and Natural Language Processing (NLP) features to generate summary related to the specified event. In NLP features, we are using Word sense disambiguation and Textual Entailment technique. Word sense disambiguation is used while calculating the weight of the word. Then we apply Textual Entailment technique on the output of the Phrase Reinforcement algorithm to find out sentences which have the strong relation between them. This is done with the help of one of the similarity measure.

The fourth step shows actual output, the i.e. summary of a specified event.

IV. RESULTS AND EVALUATION

For the result and evaluation, we are using ROUGE toolkit. ROUGE stands for Recall-Oriented Understudy for Gisting Evaluation. It includes measures to automatically determine the quality of a summary by comparing it to other (ideal) summaries created by humans. The measures count the number of overlapping units such as n-gram, word sequences, and word pairs between the computer-generated summary to be evaluated and the ideal summaries created by humans.

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 5, May 2015

To evaluate the performance of our automatic summarizer, we used the following approach. We first gathered a set of testing data by collecting the tweets for the event name specified by the user by using Twitter Stream API. The Twitter Stream API has no limit for downloading tweets at a time. In addition to this we have also taken top ten current trends from twitter and shown to the user. The user can use these current trends if he doesn't know the name of the event. We have collected more than 2000 tweets for each topic. Then we have applied filtration technique onset of tweets to remove any spam posts, short posts etc. from the same user. We also removed any non-English posts and the tweets which contain bad words. The remaining tweets are saved into a file. We then gave the files (Contains Tweets) to different human volunteers for generating manual summaries. We also gave the same set of files to our Summarizer for generating automatic summaries. We then compared the summary generated by our summarizer with the human-generated summaries.

ROUGE -N: N-gram Co-Occurrence Statistics
Formally, ROUGE -N is an n-gram recall between a candidate summary and a set of reference summaries. ROUGE -N is computed as follows:

$$ROUGE - N = \frac{\text{Count of matched n} - \text{grams}}{\text{No. of n} - \text{grams in the reference summary}}$$

In this paper for evaluation of the result, we have used ROUGE–N feature from Dragon Toolkit package. The Dragon Toolkit is a Java-based development package for academic use in information retrieval (IR) and text mining.

Document	Recall	Precision	F-Measure
CWC15	0.6680361544782252	0.8681260010678057	0.7550499187369399
AAP sweep	0.39433701657458564	0.7010435850214856	0.5047513812154696
DelhiShamedAgain	0.3953345610804174	0.7419354838709677	0.5158189827793351
Peshawar Attack	0.35576685601383057	0.7595570788294226	0.4845681607938777
Phillip Hughes	0.5	0.8606194690265486	0.6325203252032521

The above table shows the comparison of summaries generated by our Summarizer with human-generated summaries. Here the human generated summary is taken as reference summary and summary generated by our Summarizer is taken as candidate summary. The numerical data shows how automated generated summaries are related to human-generated summaries with the help of ROUGE-N technique.

V. CONCLUSION

Our conclusion is based on the result obtained with the help of ROUGE toolkit. It concludes that Phrase Reinforcement Algorithm along with two NLP features produces better summary because Natural Language Processing works at the semantic level and Phrase Reinforcement Algorithm finds important phrases from the set of tweets. In future this work can be extended by adding multimedia feature i.e. in summary along with the tweets we can also provide one option to user so that he can see the multimedia data related to that tweet for better understanding because visual (image) data is easy to understand as compared to textual. One more future enhancement in this project is summarizing user timeline.

REFERENCES

[1]. Dehong Gao, Wenjie Li, Xiaoyan Cai, Ren xian Zhang, and You Ouyang, Sequential Summarization: A Full View of Twitter Trending Topics in IEEE/ACM TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 22, NO. 2, FEBRUARY 2014.
[2]. B. Sharifi, M.-A.Hutton, and J. K. Kalita, Automatic summarization of Twitter topics in Proc. National Workshop Design Anal. Algorithms, 2010.
[3]. Gulab R. Shaikh, Digambar M. Padulkar, Template Based Abstractive Summarization of Twitter Topic with Speech Act by Asst. Prof., Department of CSE, VPCOE Baramati, Pune, India, India in June 2014.
[4]. Renxian Zhang, Wenjie Li, Dehong Gao, and You Ouyang, Automatic Twitter Topic Summarization With Speech Acts in IEEE TRANSACTIONON AUDIO, SPEECH AND LANGUAGE PROCESSING, VOL. 21, NO. 3, MARCH 2013.
[5]. Hassan Sayyadi, Matthew Hurst and Alexey Maykov, Event Detection and Tracking in Social Streams. In Proceedings of ICWSM, 2009.
[6]. J. Nichols, J. Mahmud, and C. Drews Summarizing sporting events using Twitter in Proc. IUI-12, 2012.
[7]. Joel Judd and Jugal Kalita, Better Twitter Summaries. HLT-NAACL 2013: 445-449.
[8]. Chin-Yew Lin, ROUGE: A Package for Automatic Evaluation of Summaries. In Proceedings of Workshop on Text Summarization Branches Out, Post-Conference Workshop of ACL 2004, Barcelona, Spain.
[9]. Prodromos Malak's Otis and Ion Androutsopoulos Learning Textual Entailment using SVMs and String Similarity Measures by Department of Informatics, Athens University of Economics and Business Patision 76, GR-104 34 Athens, Greece.
[10]. Xiaobin Li, Stan Szpakowicz and Stan Matwin, A WordNet-based Algorithm for Word Sense Disambiguation. In Proceedings of the 14th International Joint Conference on Artificial Intelligence.
[11]. George A. Miller (1995) WordNet: A Lexical Database for English. Communication of the ACM Vol. 38, No. 11: 39-41.
[12]. Christiane Fellbaum (1968, Ed.) WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.
[13]. https://about.twitter.com/company